Notification No: 553/2024

Date of award: 02-02-24

Name of the Scholar: Deepali Dhaka

Name of the Supervisor: Prof. Monica Mehrotra

Name of the Department/Centre: Computer Science

Topic of Research: Spammers detection in online Social Networks

Spammer detection in online social networks refers to the process of identifying and mitigating users or accounts that engage in malicious or deceptive activities, often with the intent of disseminating unwanted, inappropriate, or harmful content. Spammers exploit the interconnected nature of social platforms to spread their messages, which can include scams, phishing attempts, malicious links, and various forms of unsolicited content such as adult material or false information. The rapid expansion of online social networks has fundamentally transformed how we communicate and share information. However, this exponential growth has also ushered in a concerning issue - the rampant distribution of adult content by spammers. This problem has become increasingly apparent and intrusive, particularly with the rising prevalence of online platforms among users. It is imperative to enhance user experiences and shield individuals, especially those in younger age groups, from exposure to explicit materials.

This thesis mainly focuses on spammers distributing explicit or adult content using social networks, especially Twitter. They may target unsuspecting users, including minors, with such content. Addressing this issue is not only crucial for preserving the digital ecosystem's sanctity but also for ensuring a positive and safe user environment.

At the core of this study's solution lies a sophisticated lexicon-based methodology that revolves around the intricate interplay between users' behaviors and their core values. The central premise posits that users' actions mirror their deeply ingrained values. By deciphering these values, the model becomes adept at predicting users' tendencies and identifying those responsible for distributing adult content. A model that ingeniously amalgamates a diverse array of content-based attributes, with values occupying a pivotal position is proposed. These values are seamlessly integrated with additional features such as word entropy, lexical diversity, and context-based word embeddings. This fusion of attributes equips the model with a nuanced understanding of users' behaviors, enhancing its resilience and accuracy in categorization. The analysis quantifies the discriminative power of these attributes, unveiling the complex interplay between them and highlighting their cumulative impact on the model's performance. The findings underscore the indispensable contribution of attributes like values, word entropy, lexical diversity, and contextual embeddings in shaping the model's success.

These kinds of spammers have been studied and detected using various hybrid features and machine-learning approaches in the past. To have greater insight into data prevailing in the form of text on platforms like Twitter, their correct vector representation is paramount. Ourgoal is to understand what encoding techniques are more suitable for representing long text documents. Therefore, anovel deep learning model is proposed consisting of a Universal Sentence Encoder (USE) as a feature extractor and an artificial neural network (ANN) as a classifier. All the sentence vectors representing the tweets of a user are transformes into a document vector. These vectors are used as high-quality features to be processed by the artificial neural network for classification. To check the effectiveness of our proposed model, different sentence embedding techniques such as Doc2Vec, Infersent, and Sentence-Bert have been used and compared with the proposed model. Experimental results show that the proposed model outperforms all of them. Our results show that a simple ANN combined with a USE-based deep learning approach can be a robust solution for the detection of spammers on Twitter.

Moreover, conventional approaches to spam detection have predominantly concentrated on identifying spam within individual platforms. However, this thesis ventures beyond the limitations of single-platform analysis. It undertakes a comprehensive exploration encompassing cross-domain spam detection techniques applicable to email and web spam, social spam, and opinion spam. By engaging in meticulous comparisons, this study seeks to unravel diverse challenges in this domain. Significantly, it pioneers an unprecedented endeavor, marking the first in-depth literature examination within the realm of cross-domain spam detection.

In a groundbreaking move to substantiate these claims empirically, a novel SBERT-based model is introduced. This innovative model is ingeniously designed to target spammers disseminating inappropriate content on Twitter, capitalizing on data derived from Reddit. The inherent challenge of bridging the semantic gap between these distinct platforms is met through the utilization of the valueDict lexicon—a feature engineering concept previously proposed. The incorporation of these shared features culminates in a rich dataset, subsequently employed in a machine learning algorithm to assess the model's performance across varied scenarios, including within-domain, mixed-domain, and cross-domain contexts.

Crucially, the proposed model's versatility and adaptability are showcased through its application to benchmark datasets.

This holistic approach marks a pivotal shift in the domain of spammer detection, recognizing the complex interplay between different online platforms and their shared characteristics. By unifying insights from different domains, the thesis stands as a testament to the evolving landscape of digital communication. The SBERT-based model emerges as a beacon of innovation, ushering in a new era of cross-domain spammer detection that transcends the confines of singular platforms and fosters a safer and more secure online ecosystem.